Mapping The Genetic Relationships of the World's Languages

Dr. Stephen Huffman

glbh@msn.com

[Editor's note: This paper describes a series of maps produced by Dr. Huffman using a pre-release version of GMI's World Language Mapping System (WLMS) GIS data set. Most of these maps have been adjusted to use the current released version of the WLMS along with GMI's Seamless Digital Chart of the World. Images of the maps, Portable Document File (PDF) files of the maps suitable for large-format printing, and Dr. Huffman's data table and ArcGIS project files, are available in Dr. Huffman's user area on the World Language Mapping System website at http://www.gmi.org/wlms/users/huffman.]

The Genetic Classification of Languages

Classification is the art of grouping things together according to some set of criteria. We all use classification all the time to understand and interact with the world around us. We talk about 'cars' (a particular class of vehicles that are neither trucks, busses, trains, planes, motorcycles, etc.), 'vegetables' (a specific group of edible plants), 'clothes' (certain kinds of things we wear on our bodies, as opposed to hats, jewelry, etc.), and so on. A classification based on a useful set of criteria enables us to talk easily about related things, and often leads to useful insights concerning the relationships among items within a class.

Languages also can be classified, and the criteria used to group them can help us to understand how languages relate to one another in various ways. There are a great many ways to classify languages. The simplest is to simply group them together based on where they are spoken. This way, we can talk about the languages of Africa or Asia, or France. This type of classification is sometimes used for areas where there are a large number of poorly understood languages, such as in the Caucasus, where classifications based on deeper understandings of the languages can be controversial. It can also be used to see where there are pockets of people belonging to particular ethnic groups (speaking their own languages) within some larger region.

Languages can also be classified according to features of the languages. For instance, languages can be classified according to whether or not tones are used to distinguish meaning among words, and if so, by how many and what sort of tones are used. Or, they can be classified according to their sound systems; for instance, how many

vowels or consonants they utilize, or whether they use specific sounds, such as the clicks found in some African languages. Languages can be classified by how the words of a typical sentence, for instance the Subject (S), Object (O), and Verb (V), are ordered. In this scheme, English would be characterized as an "SVO" language.

Obviously, there are a nearly endless number of ways to classify languages. But one of the most fascinating ways to classify them is genetically. This form of classification attempts to group together languages that have descended from a common ancestral tongue.

Consider the French, Italian, Spanish, Portuguese and Romanian languages. All of these developed out of the ancient Latin language. During the course of its history, the Roman empire conquered and colonized much of Europe. Latin-speaking soldiers, merchants, officials, and common people settled throughout the empire, bringing their Latin tongue with them. And eventually, the local, non-Latin speaking peoples adopted Latin themselves as they embraced Roman culture. As long as the empire existed, with its extensive road system and easy movement throughout, the Latin spoken by the people throughout the empire was probably fairly similar. But after the fall of the Roman empire, when long-distant movement become much more difficult and dangerous, the speech of the Latin-speaking peoples of France, Italy, Spain, Portugal and Romania began to drift apart. Fairly modest differences in accent and vocabulary, such as today characterizes the difference between American and British English, eventually developed into distinctions of vocabulary, pronunciation and even syntax (sometimes under the influence of neighboring, non-Latin languages) that were so extensive that the different ways of speech evolved into new and different languages. Even so, it is clear that all these languages are very similar to each other. They are similar because they each developed out of the same "proto-language," which was Latin. The languages that are directly descended from the language of the Romans therefore are genetically related, and are called Romance languages (from "Roman"). Each Romance language is more closely related to the other Romance languages than it is to a non-Romance language such as German or English.

German, meanwhile, is clearly more similar to languages like Swedish, Danish, Icelandic, Dutch, and English than it is to any of the Romance languages. By studying

the systematic nature of their similarities, scholars have deduced the existence of a language they refer to as ProtoGermanic, which must have been spoken by the ancestors of the Swedes, Germans, English, and other related peoples. ProtoGermanic must have been spoken before the art of writing reached those peoples. Even so, scholars have been able to reconstruct a substantial portion of the ProtoGermanic language by studying the languages that descended from it. From this, and other clues, scholars believe ProtoGermanic was probably spoken about 2500 years ago.

But genetic classification of languages can reach back even further in time. The Germanic and Romance languages are, in fact, more closely related to each other than they are to most of the other languages of the world. This was recognized as far back as the 1700s. In fact, the Germanic and Romance languages are just two branches of a much larger family of languages known as IndoEuropean. The IndoEuropean languages include most of the languages of Europe (with the exception of Basque, Hungarian, Finnish, Estonian, and the Saami languages of the far north), as well as many of the languages of India, Iran, and Afghanistan. This family includes roughly 200 living languages.

Scholars demonstrated the relationship of the IndoEuropean languages by first comparing basic words in many of the languages of Europe and India. After analyzing their similarities, they were able to tentatively reconstruct what may have been the original form of many words in what scholars call the ProtoIndoEuropean language. For example, consider the following set of basic words taken from several branches of the IndoEuropean language family:

	Meaning			
Language	father	mother	two	three
AngloSaxon	faether	modor	twa	thrie
Latin	pater	mater	duos	tres
Greek	pater	meter	duos	tri
Sanskrit	pitar	matar	dva	trayas

By studying these words, scholars reconstructed the ProtoIndoEuropean forms for the words as follows (the asterisk preceeding each word means it has been reconstructed): *pater, ‘father’, *mater, ‘mother’, *duwos, ‘two’, *treyes, ‘three’. In like manner,

scholars examined the grammar of the various IndoEuropean languages, and have attempted to reconstruct some of the grammar of ProtoIndoEuropean.

Over the past 150 years, linguists have demonstrated that nearly all the languages of the world belong to some language family, each of which is descended from some now-lost proto-language (those languages which have not been shown to be related to other languages are called isolates). Among these language families are the SinoTibetan; which includes Chinese and Tibetan, NigerCongo; which includes many of the languages of sub-Saharan Africa; Austronesian, encompassing many of the languages of Oceania; and AfroAsiatic, which includes many well known middle eastern languages such as Hebrew and Arabic.

But the art of genetic classification is not without controversy. One of the fiercest debate in linguistics today concerns "high-level" classifications: the grouping of generally accepted language families (such as IndoEuropean) into still higher-level groups. This is because as linguists group higher-level language families together, the relationships that are revealed necessarily extend further back in time. For instance, as mentioned above, most linguists believe that ProtoGermanic was spoken in a region of northern Europe about 2500 years ago. But ProtoGermanic itself developed out of the ProtoIndoEuropean language that must have been spoken much earlier. Just how much earlier is a matter of great debate; estimates range from about 5000 years ago to over 10,000 years ago.

All linguists accept the validity of the IndoEuropean language family. But some historical linguists believe that IndoEuropean itself can be shown to be related to other language families, in a higher-level grouping known variously as Eurasiatic or Nostratic (the two names refer to slightly different groupings of language families, arrived at by different techniques). If valid, the age of these higher-level language groups must be at least 10,000 years old. Other, more conservative, linguists believe that after so much time, languages have diverged so much from both their ancestral and sibling languages that it is impossible to demonstrate relationships among them.

Still, a very few linguists even believe that it is possible to recover a dim glimpse of a time when the first human language was spoken. They believe that it is possible to reconstruct at least a few words of such a language. Other linguists argue that such

ancient words are, and will always remain, unrecoverable. It will be a long time, if ever, before this debate is resolved.

In any case, it is clear that language families provide unique evidence of the relationships among groups of people far into the past – much further than any detailed recorded history. And the higher-level language groups obviously indicate relationships further back in time.

I should note that the debate over the time-depths at which linguistic relationships can be demonstrated is sometimes carried on with a level of vitriol that is quite excessive. To even allow for the possibility of finding remote linguistic relationships (at least more remote than the fairly obvious ones) is to open oneself up to attack by some scholars. This is because the debate over language relationships has implications for historical studies, anthropology, and the proper methods of linguistic investigation. There are some scholars who are extremely upset when it is suggested that their standard of proof may not be the only correct one. I am not trying to take sides, just to caution readers that some linguists take great umbrage over some of the high-level classifications used in some of these maps. However, most of the language families presented in most of these maps are uncontroversial.

The language classifications used on these maps derive primarily from two sources. The lower-level language groupings are taken mostly unchanged from the *Ethnologue* classification (14th Edition). For higher-level language groups, I have relied on Merritt Ruhlen's A Guide to the World's Languages (published 1987, 1991 by Stanford University Press). This book provides an excellent overview of the current status of the genetic classification of the languages of the world, along with good discussions of the history of those classifications and the debates over the high-level language classifications. I used Ruhlen's work to modify the language family field (field G) in the *Ethnologue* attribute table of the language polygons supplied with the World Language Mapping System (www.gmi.org/wlms)

The Maps

These maps all show the distribution of languages of the world, to varying degrees of detail, based on their genetic relationships. The one region of the world that is

not well represented here is the Americas, because the distribution of indigenous languages is rather sparse. At a later time, I hope to create a set of maps that focus on those areas where indigenous languages are still spoken widely in the Western Hemisphere.

The map "The World's Language Phyla" is the one most likely to draw condemnation from more traditional historical linguists. It shows the worlds languages, grouped according to the highest-level classifications that have been proposed by linguists. Many of the proposed groupings shown in this map are still highly controversial. The language families in this map are discussed below.

Undoubtedly the greatest classifier of languages in history was Joseph Greenberg. His classification of the languages of Africa in the 1950s revolutionized the field of African linguistics, and is the foundation for all work on classification in Africa today. During his life, Greenberg studied languages throughout the world, and proposed many high-level classifications, some of which are still quite controversial. However, after studying his work and the responses to it, my personal opinion is that most of his proposals will stand the test of time. Greenberg had an extraordinary grasp of the world's languages, and an incredible talent for synthesizing enormous amounts of information. Several of the groupings presented in this map reflect Greenberg's work.

Certainly one of Greenberg's most controversial classifications was that of the languages of the Americas. In 1987, Greenberg proposed that all the indigenous languages of the Americas belonged to one of three families, while more traditional linguists group these into two hundred or more language families. According to Greenberg, most languages in the Americas belong to a large family he calls Amerind. Those languages related to Navajo and Apache are designated the NaDene language family, and the remainder belong to the EskimoAleut family. Linguists are highly polarized on this issue (to say the least), but I believe it is more likely that Greenberg's classification is essentially correct. Note, however, that most specialists in American Indian languages still reject Greenberg's Amerind language family.

Another high-level language family proposed by Greenberg is the IndoPacific language family. A great deal more research needs to be done on the languages of this region of the world, and so this is a more tentative hypothesis. Basically, he has

combined all the non-Austronesian languages of Papua New Guinea into a single language family, as well as some languages in islands in the vicinity of New Guinea. Most researchers think Greenberg's classification is too broad, and attribute similarities among these languages to widely scattered remnants of more ancient (and perhaps now lost and forgotten) languages.

The Australian language family is fairly non-controversial, though at least one prominent Australian linguist (Dixon) believes that the languages of Australia are similar, not due to a common ancestor, but because they have been in contact so long that they have all borrowed extensively from each other.

Eurasiatic is another of Greenberg's proposed high-level language families. For at least two hundred years, linguists have noted similarities between IndoEuropean and other language families, but it was not until the mid 1900s that systematic work was done to try to determine exactly which language families might be related. The Russians worked particularly hard on this problem, and have proposed a group of related language families (one of which is IndoEuropean) that they refer to as Nostratic. Using different methodologies, Greenberg and the Russians have come up with high-level language groupings that are fairly similar. This map follows Greenberg's proposal, which I believe is the more likely one, and the one that the Nostratic hypothesis seems to be moving towards.

The proposed Austric language family consists of the concatenation of four lower level language families – the Austronesian language family, the Austroasiatic, Miao-Yao, and Daic language families. Each of these lower level groups is relatively uncontroversial. This combination is much more so.

Probably the most controversial of all proposed language families in this map is the DeneCauscasian language family. This was not one of Greenberg's proposals, but one put forth by a handful of historical linguists. DeneCaucasian combines the Sino-Tibetan, Basque, NaDene, and North Caucasian language families along with a few small languages scattered throughout Asia. If this is valid, and only a few linguists currently defend it, it would represent a very old language family that was swamped by speakers of languages from other language families.

The remaining language families depicted on this map are generally widely accepted. All the indigenous languages of Africa belong to four language families: AfroAsiatic (containing such well known languages as Ancient Egyptian, Hebrew and Arabic), NigerCongo (whose best known subgroups are the Bantu languages which cover most of southern Africa), the little-known NiloSahran language family (about which some linguists are skeptical), and the Khoisan language family, whose members are best known to linguists for their distinctive click sounds. A number of linguists are beginning to consider combining the NigerCongo and NiloSaharan languages into a larger grouping, but this idea is still very tentative.

The Kartvelian language family consists of a few very closely related languages spoken in the country of Georgia, just south of the Caucasus mountains. These languages are not related to the North Caucasian languages, spoken on the other side of the Caucasus.

Dravidian is language family that is concentrated in southern India.

Most of the remaining maps show the languages in particular areas of the world. By studying these maps, one can see how the major language families are distributed throughout the world.

The “IndoEuropean Language Family” map shows the distribution of the IndoEuropean language family.

The “Eurasiatic Language Family” map is quite similar to the IndoEuropean language map, but it shows the distribution of the proposed Eurasiatic language family, of which the IndoEuropean language family is one branch.

The “Indigenous Languages of Africa” map shows the distribution of the four major language families of Africa as demonstrated by Greenberg.

The “Horn of Africa” map depicts, not surprisingly, the languages of the Horn of Africa. The language families are all represented by shades of a particular color, with each language delineated by its own color.

The “Languages of the Sudan” map is organized on very similar principles to “Horn of Africa” map. It merely shows the languages in Sudan a bit more closely. This map also emphasizes how much of the Sudan is essentially uninhabited.

The “Languages of Southeast Asia” map shows how the different language families of Southeast Asia are intermixed in the region.

The “Languages of the Indian Subcontinent” highlights the relationship of language families in India.

The “Languages of China” shows the languages in the region of China, and the “Languages of the Philippines” details the languages of the Philippines.